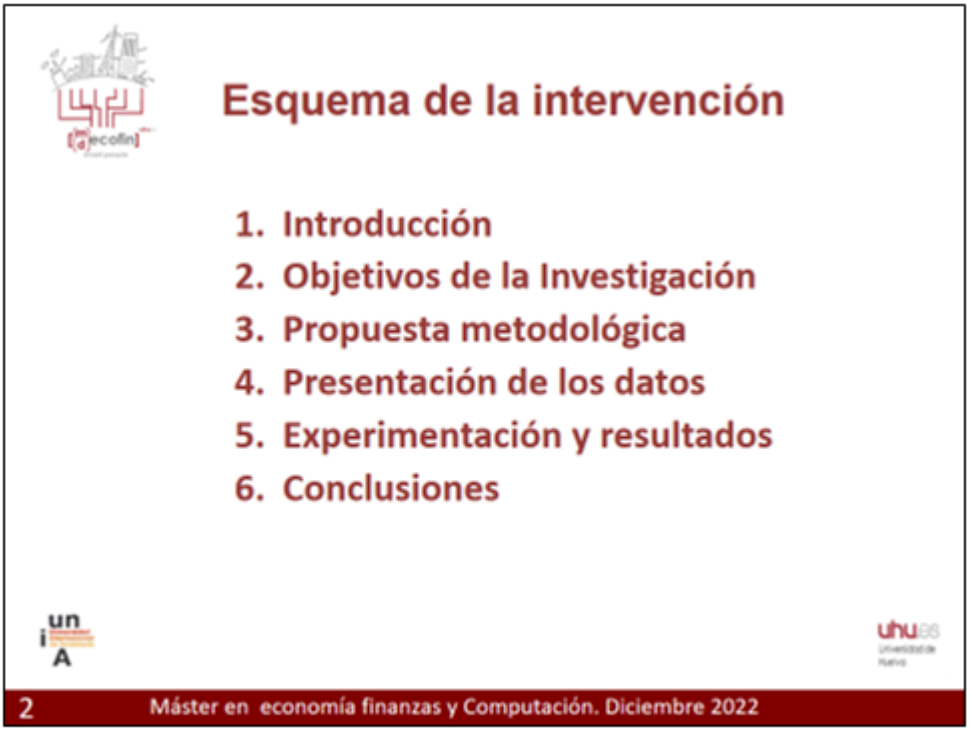
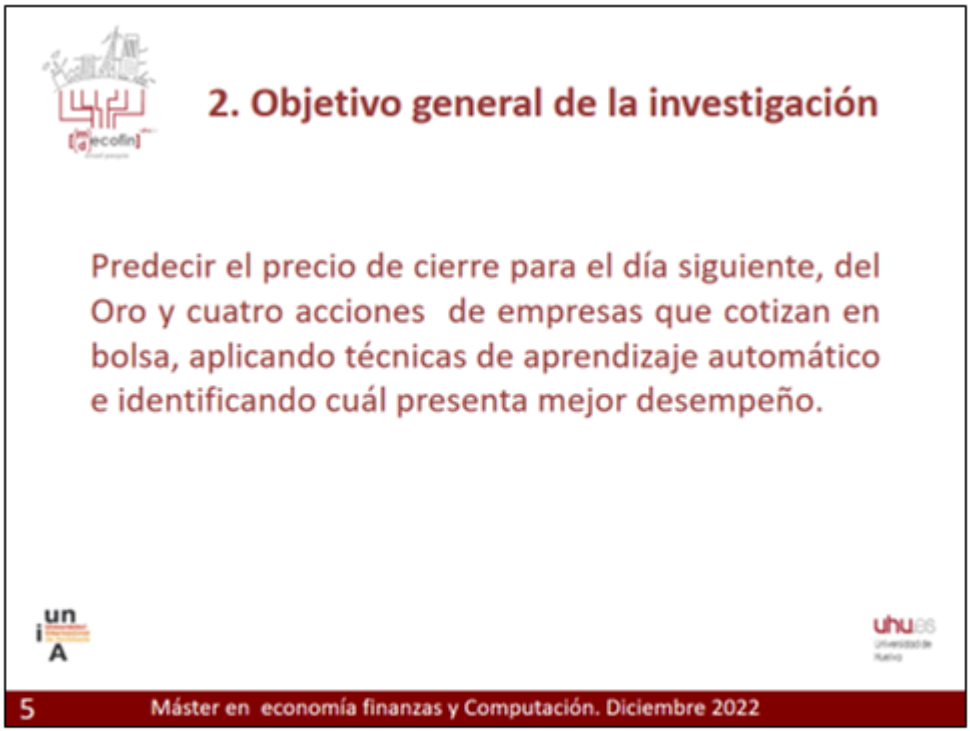
Buenos días el título de este trabajo de fin de máster es:
Predicción de precios de acciones mediante modelos de aprendizaje automático
Del autor Yunio Seijo Manso, bajo la supervisión de Diego Marín Santos y Manuel Emilio Fernández Arias



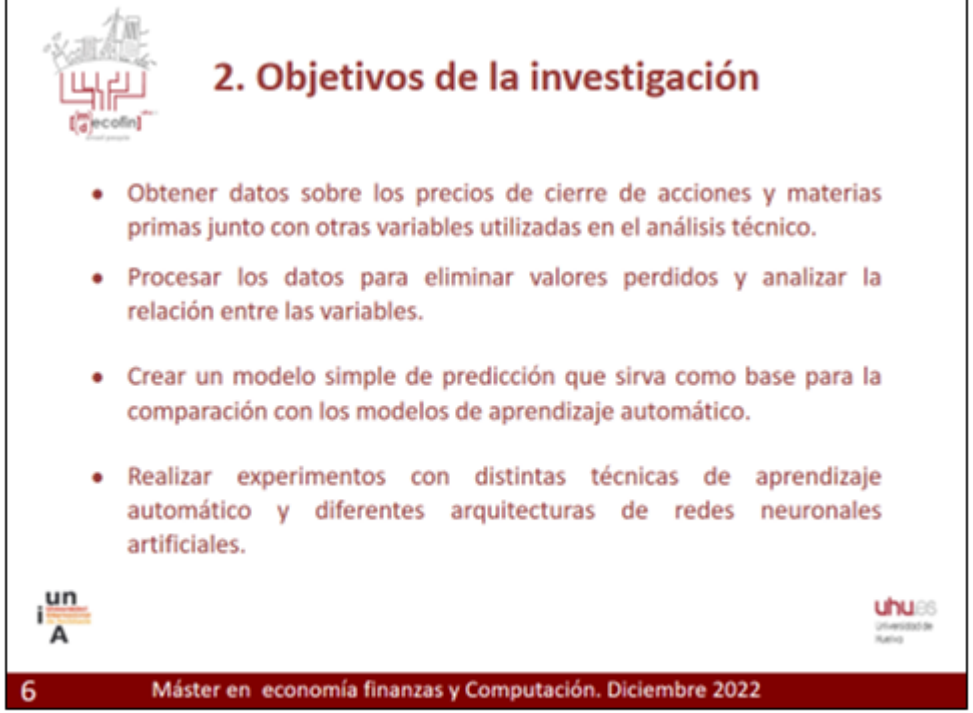
El esquema de mi presentación es…

Existen diversos debates sobre la predicción de los rendimientos de los mercados. Eugene Fama en 1970 introdujo la hipótesis del mercado eficiente, donde establece que el precio actual del activo refleja siembre toda la información previa. Por otra parte Gordon Malkiel en 1999 planteo la hipótesis de la marcha aleatoria, que afirma que el precio de una acción cambia independientemente de su historia, es decir, que el precio de mañana sólo dependerá de información de mañana, independientemente del precio de hoy.
Por otro lado, Craig Mackinley afirma que, los precios de las acciones pueden predecirse al menos en cierta medida.

La inteligencia artificial (IA) en las finanzas se vuelve un tema cada vez más popular. Numerosos estudios han sido publicados dando como resultado varios modelos e interpretaciones de cómo hacer predicciones.
Según el instituto Alan Turin se estima que entre el 60 y el 73% de las operaciones diaria son llevadas a cabo por algún tipo de algoritmo automatizado.



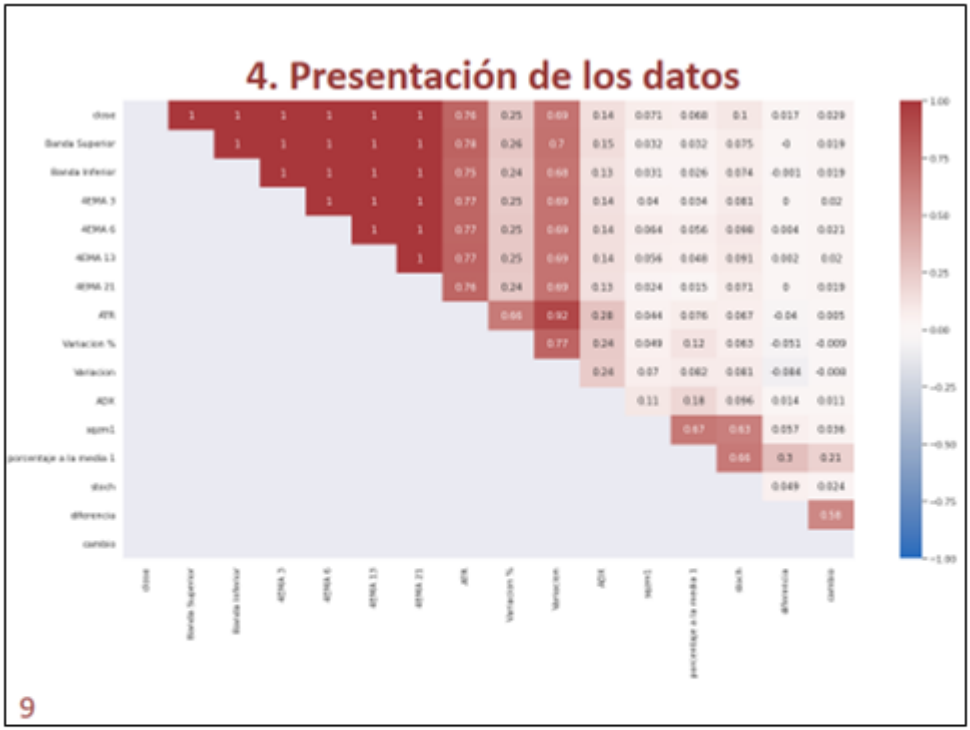
Las contradicciones entre diferentes autores y enfoques referentes a la predicción de precios de acciones en el mercado de valores, ha motivado esta investigación, como objetivo se propuso.



Y se plantean los siguientes objetivos específicos.

La Propuesta metodológica para este trabajo sería entonces partiendo de los datos extraídos de la plataforma TradingView.
Revisar y procesar los datos.
Seleccionar cuáles serían las técnicas de regresión con las que haría las predicciones.
Extraer las características que pudieran representar mejor el modelo.
Implementar los modelos y entrenarlos.
Y por último analizar y comparar los resultados obtenidos.

En este trabajo los activos analizados fueron.
Cada base de dato contenía 28 variables utilizadas en el análisis técnico, el periodo analizado va desde enero de 1980 hasta el 10 de octubre del 2022. Se utilizaron diferentes métodos para la selección de características y se crean 5 conjuntos de datos, en la presentación se hará referencia al conjunto uno, que consiste en 16 variables con 3 retrasos de cada una dando como resultado 48 variables que se usan como variables independientes.



Para la selección de las variables independientes se utilizó inicialmente como método de filtrado, la correlación entre las variables utilizando el método de pearson. valores cercanos a 1 indican la más alta correlación positiva, los cercanos a cero no tienen correlación y -1 sería la más alta correlación negativa. En la escala de colores, el rojo representa una alta correlación y el azul alta correlación negativa.
En este caso notar que las bandas bollinger superior e inferior y las medias móviles exponenciales tienen una altísima correlación con la variable precio de cierre (close).

4. Presentación de los datos

Estadística descriptiva

	count	mean	std	min	25%	50%	75%	max
close	10715	749.416	506.182	252.1	364.65	435	1224.05	2063.564
media de Bollinger	10715	748.315	505.124	254.871	364.257	434.005	1224.071	1972.063
Banda Superior	10715	767.808	517.532	256.558	371.881	446.338	1250.601	2053.162
Banda Inferior	10715	728.823	493.111	238.462	356.445	422.748	1200.515	1925.393
4EMA 3	10715	748.294	504.973	255.389	364.211	433.063	1224.107	1950.632
4EMA 6	10715	749.148	505.855	253.971	365.223	434.667	1224.067	2021.033
4EMA 13	10715	748.925	505.615	254.437	364.742	434.417	1224.272	1997.441
4EMA 21	10715	746.199	502.957	257.584	365.145	433.383	1219.311	1922.828
ATR	10715	10.786	10.34	0.29	2.975	6.874	16.215	94.509
Variacion (%)	10715	1.152	1.105	0	0.404	0.978	1.584	15.287

10

Como parte de la metodología se realiza un análisis descriptivo de los datos, con el objetivo de conocer los rangos mínimos y máximo de cada una, conocer sus medias y detectar posibles valores anómalos como pudieran ser precios negativos. etc

5. Experimentación y resultados

Técnicas de predicción utilizadas

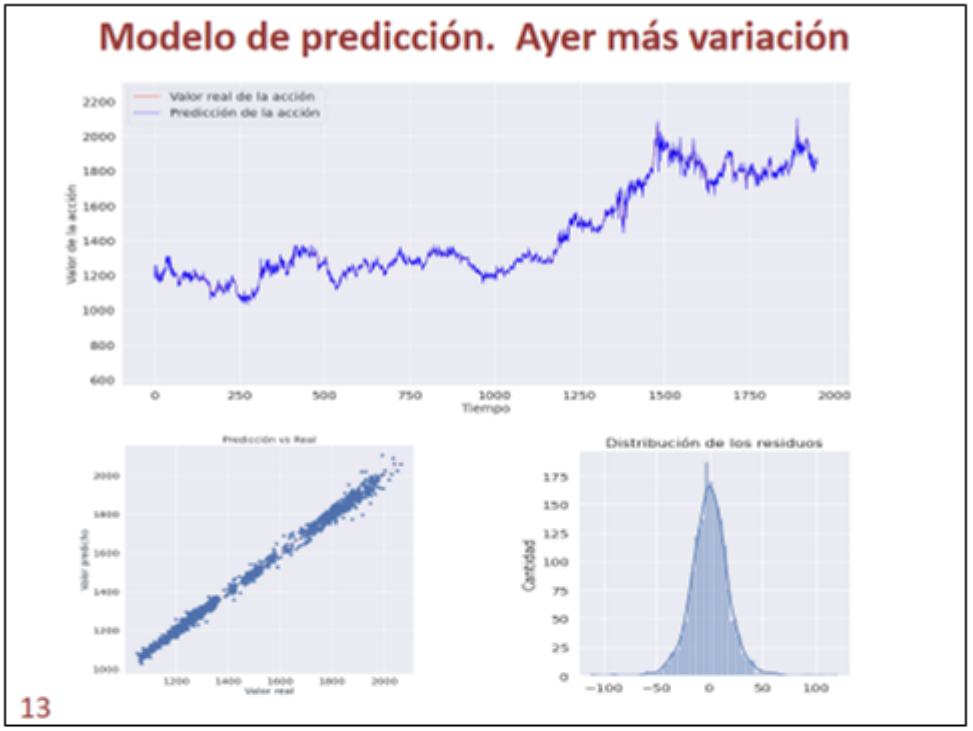
- Ayer más variación
- Mañana igual que ayer
- Random forest
- Regresión lineal
- Regresión lineal 30-1
- Regresión lineal 200-1
- Red neuronal convolucional
- Red neuronal densa
- Red neuronal LSTM
- Red neuronal BLSTM

Luego del análisis exploratorio inicial y la selección de características , se decide implementar varias técnicas,
Se comienza con técnicas básicas que serían: ayer más variación y mañana igual que ayer.
Luego técnicas más complejas como Random Forest
Regresión lineal y de esta dos variantes, con diferencia en su forma de entrenamiento.
Además se utilizaron algoritmos de redes neuronales con diferentes arquitecturas como la red neuronal convolucional, una red densa, una red recurrente LSTM de su nombre en inglés Long short term memory y una red neuronal bidireccional LSTM.

Seleccionar los datos de aprendizaje y entrenamiento es fundamental en el proceso de experimentación de los modelos de aprendizaje automático, posteriormente es necesario la validación de lo aprendido por estos modelos, lo cual se realiza sobre otra parte de los datos en nuestro caso sobre un periodo futuro.

Las series temporales tienen el problema que no se pueden entrenar con datos en el futuro y validar con el pasado porque precisamente lo que se busca es la Acción de anunciar un hecho futuro.

Entonces tenemos que entrenar con el pasado, en este caso lo hago desde 1980 hasta aproximadamente 2015 y se valida desde el 2015 hasta el 2022, esto tiene como desventaja que estaría prediciendo basado en precios pasados muy alejados en el tiempo.



A continuación se analizan los resultados obtenidos por cada uno de los modelos planteados anteriormente, en este caso sobre la serie temporal de los precios del oro.

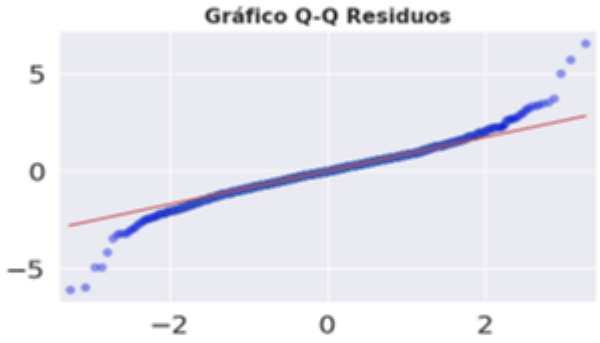
En el modelo básico ayer más variación podemos ver que los resultados son bastante buenos aparentemente, en la grafica superior podemos observar que los precios reales en color rojo y los predichos en color azul, están superpuestos, semejando así una predicción perfecta.

Mirando la gráfica del valor real contra el valor predicho en la imagen inferior izquierda se puede decir que los valores predichos se acercan bastante a una función lineal con Y=X cumpliendo así el supuesto de linealidad.

La ultima grafica representa un histograma donde se refleja la distribución de los residuos, aquí se observan que estos parecen seguir una distribución normal sin embargo.

Modelo de predicción. Ayer más variación

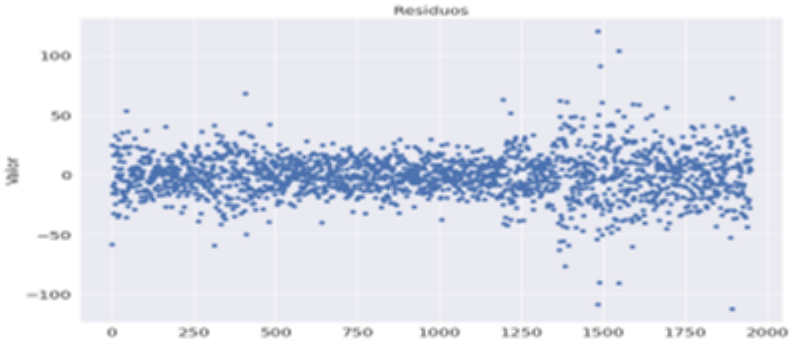
	Estadístico	p_value	Normalidad
K-squared	184.463	8,80E-38	Falso
Shapiro-Wilk	0.965	2,35E-18	Falso



14

después de aplicar los test correspondientes se rechaza la hipótesis de normalidad de los residuos. Los test aplicados fueron K-squared y Shapiro-Wilk.
Utilizando el método gráfico de cuantiles se puede comprobar que los residuos no siguen una distribución normal aunque en el gráfico anterior de distribución parecía la contrario

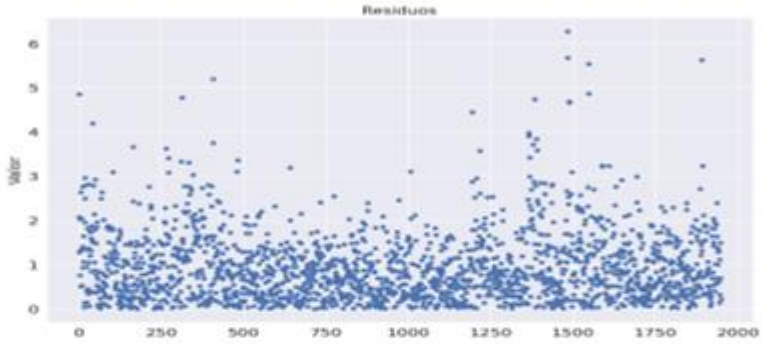
Modelo de predicción. Ayer más variación



15

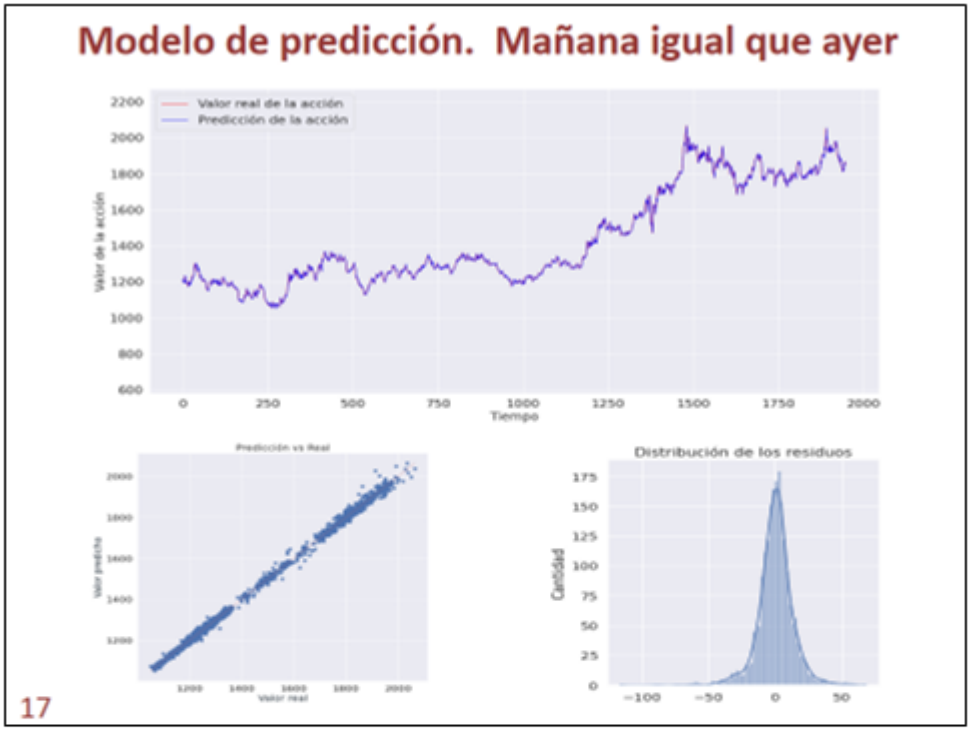
Otra análisis realizado fue el de homocedasticidad y en el gráfico se muestran como los residuos no son homocedásticos, destacar que en la serie de precios va aumentando en valor por lo que 20 de dólares de error cuando el precio es de 1050 dólares que es el mínimo para ese periodo representa mucho más que cuando el precio es de 2063 dólares por lo que se decide analizar también el error en término porcentual como se muestra en la figura sig

Modelo de predicción. Ayer más variación



16

Y en términos porcentuales los errores tampoco son homocedásticos.



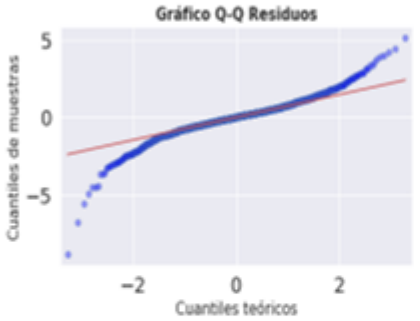
En el modelo básico mañana igual que ayer podemos ver que los resultados son bastante buenos, en la grafica se puede observar que los precios reales y los predichos están solapados.

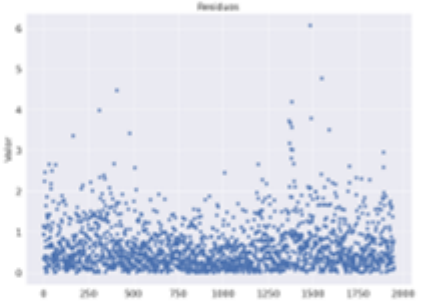
Mirando la gráfica del valor real contra el valor predicho en la imagen inferior izquierda se puede decir que los valores predichos se acercan bastante a una función lineal con Y=X.

La grafica inferior derecha representa un histograma con la distribución de los residuos, aquí se observan que estos se asemejan a una distribución normal y otra vez más luego de realizar los test se rechaza la hipótesis de normalidad de los residuos.

Modelo de predicción. Mañana igual que ayer

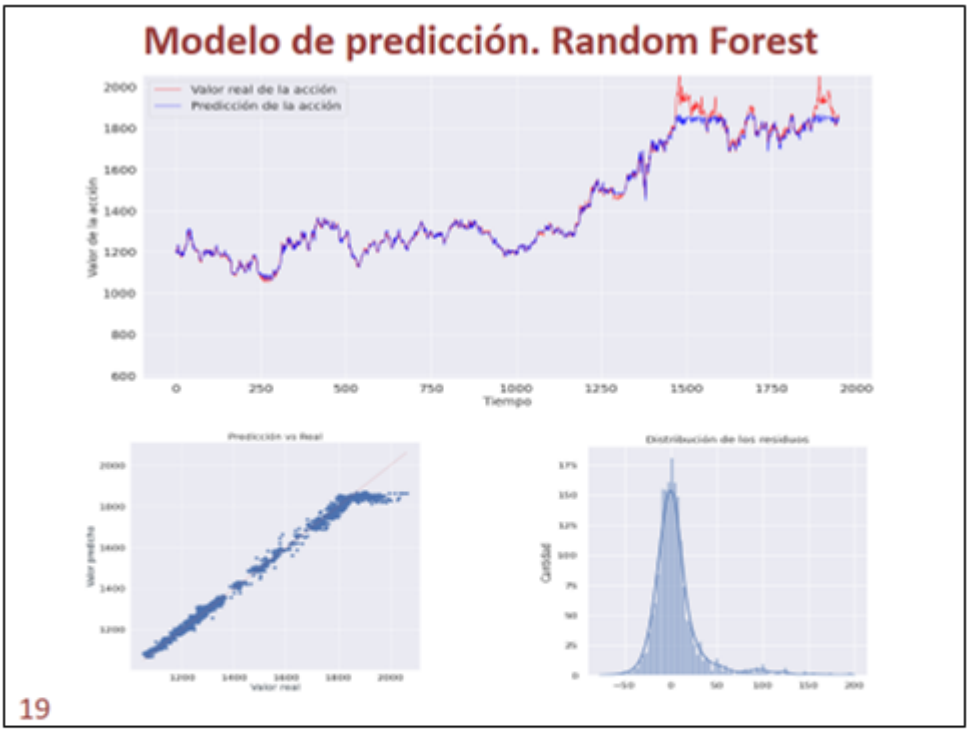
	Estadístico	p_value	Normalidad
K-squared	406.676	4.914e-89	Falso
Shapiro-Wilk	0.933	9.574e-29	Falso



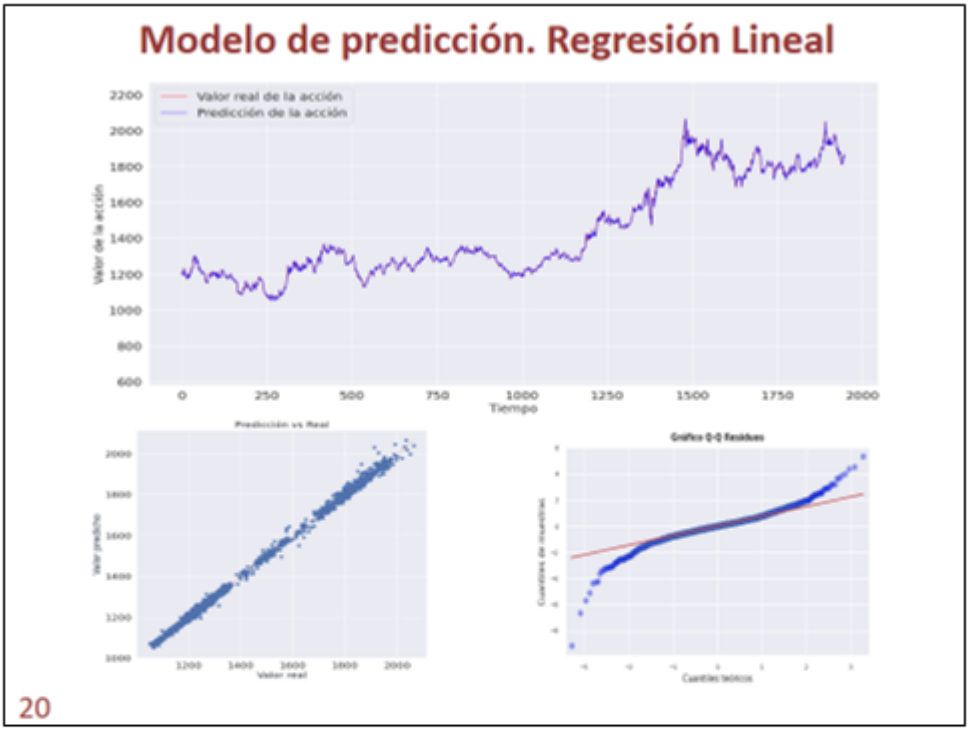


18

Los test aplicados nuevamente fueron K-squared y Shapiro-Wilk.
Utilizando el método gráfico de cuantiles se puede comprobar que los residuos no siguen una distribución normal aunque en el gráfico anterior de distribución parecía la contrario
Otro análisis importante sería el de igualdad de varianza de los errores a lo largo del tiempo y en la figura inferior derecha se puede ver que estos no son homecedásticos.



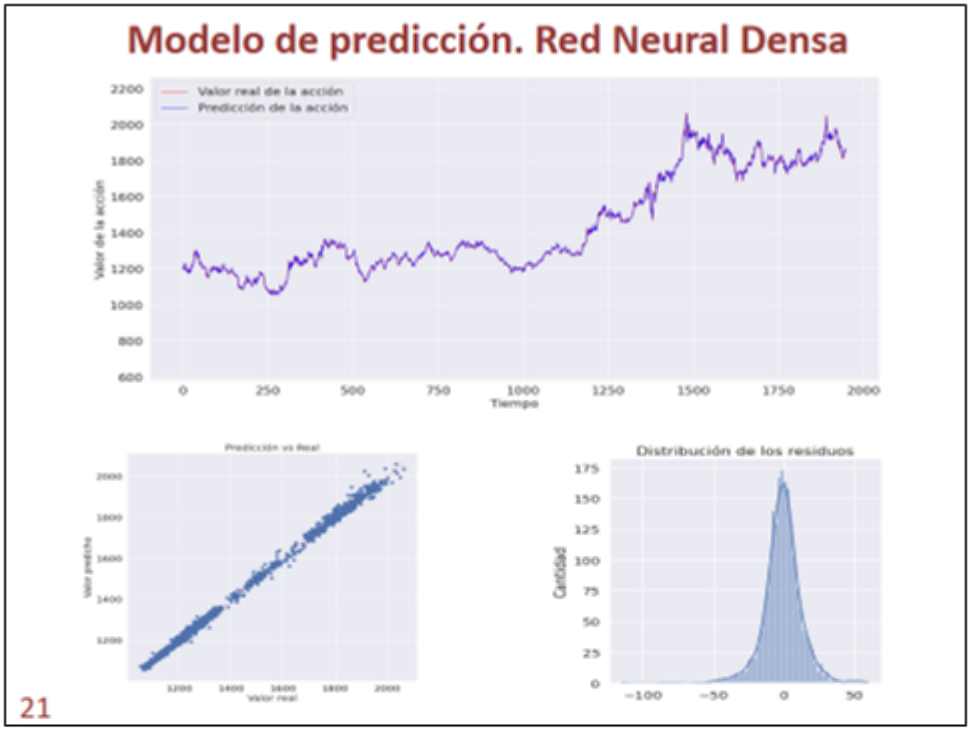
Estos serían los resultados aplicando el modelo de random Forest dónde podemos ver que en validación el algoritmo no pudo ser capaz de predecir con una mejor exactitud valores que no había visto en la etapa de entrenamiento o sea los valores por encima de los 1850 dólares aproximadamente.
Este es uno de los problemas fundamentales que tienen los árboles de decisión y es que normalmente tienden a sobre aprender



El modelo de regresión lineal también parece dar una buena aproximación al valor real, visualmente aquí podemos decir que casi que están solapados el precio real y el precio predicho.

En la gráfica del precio real versus el precio predicho que se muestra en la figura inferior izquierda viendo….¿ cómo se cumple el supuesto de linealidad? se acerca bastante a la función lineal esperada y

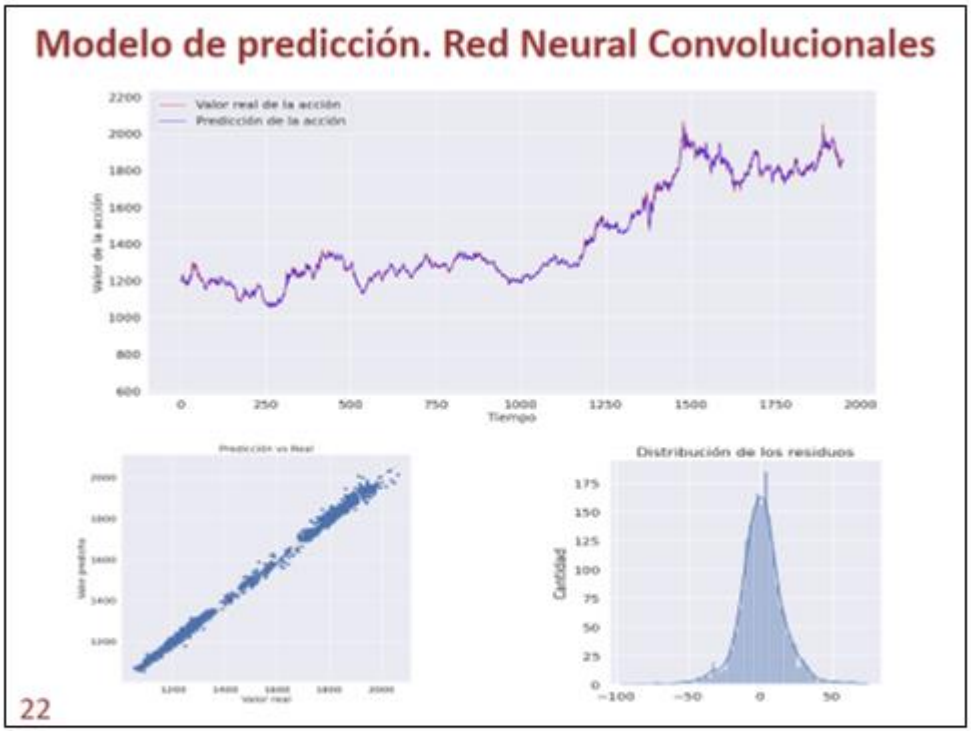
la distribución de los errores utilizando el método gráfico de cuantiles se puede ver que no siguen una distribución normal pues los puntos se alejan de la línea de referencia esperada.



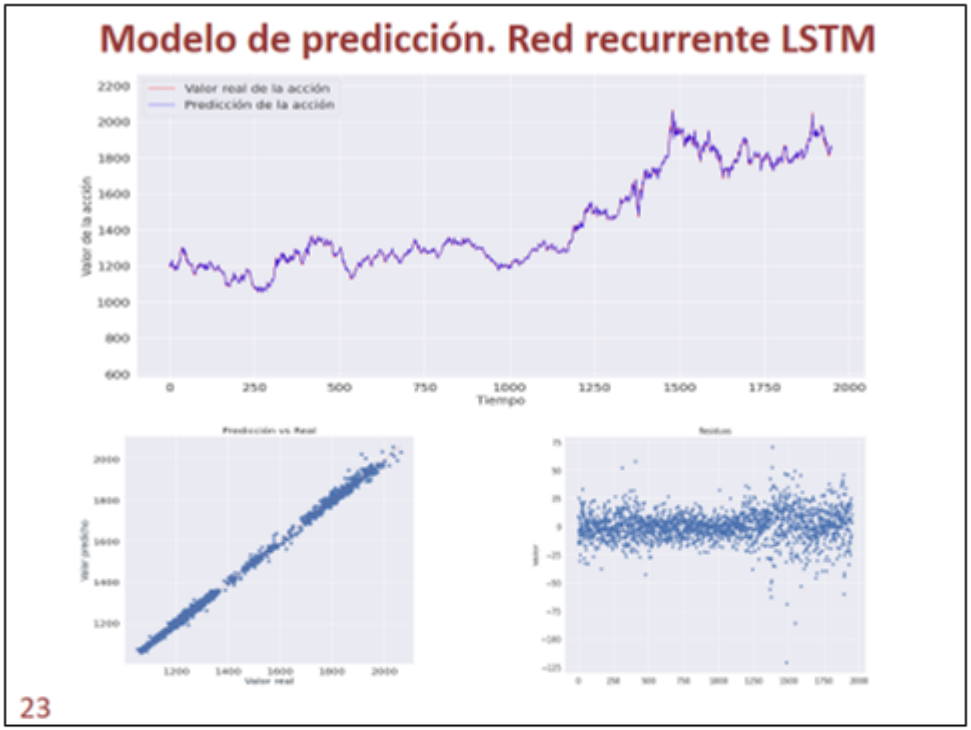
Como había comentado se utilizaron 10 métodos para predecir el precio y evaluarlos entre ellos, pero en el caso de las redes neuronales utilizando la misma arquitectura y cambiando parámetros como el tamaño del batch, la ratio de aprendizaje o cuántas neuronas tendrían por capa, se probaron hasta 54 modelos para esta arquitectura y se selecciona la de menor error cuadrático medio.

Aquí se puede ver que esta red, aún con valores que no había visto antes, hace una aproximación cercana al precio real.

En la gráfica de la distribución de los errores se puede ver que realmente la media no está en cero , sino que está un poquitico desplazado hacia los valores negativos por lo que se puede decir que los valores predichos fueron un poco mayor al precio real, desde el punto de vista de la linealidad se puede decir que se aproxima al función esperada.

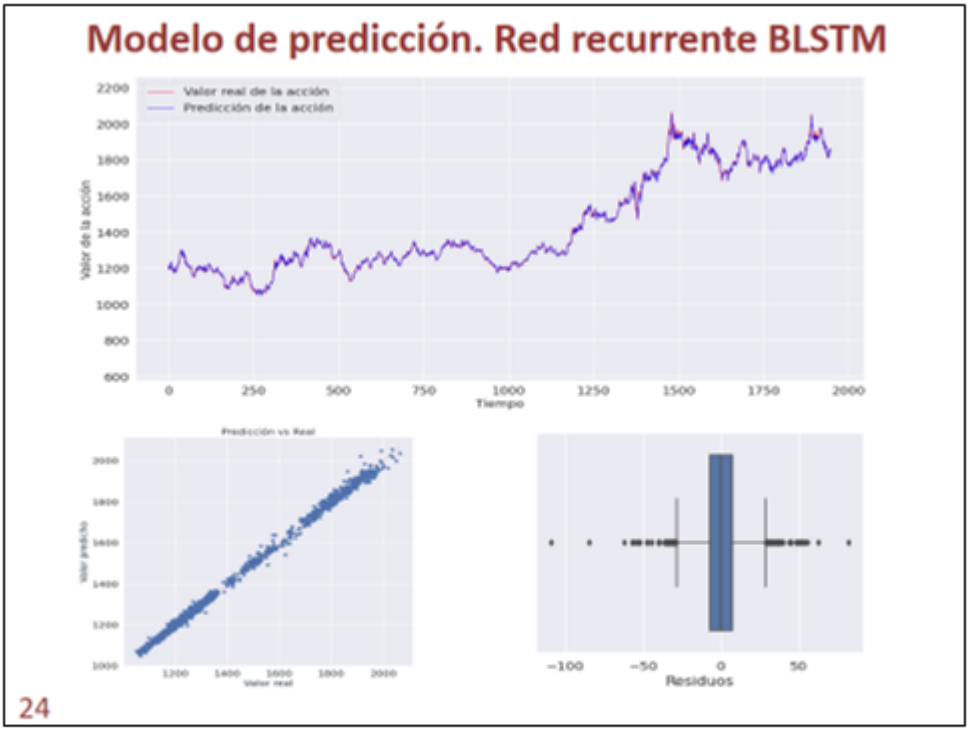


El modelo convolucional también tuvo una predicción cercana a los valores reales, como se puede ver en la figura.
Y los resultados son semejantes a los realizados por los modelos anteriores

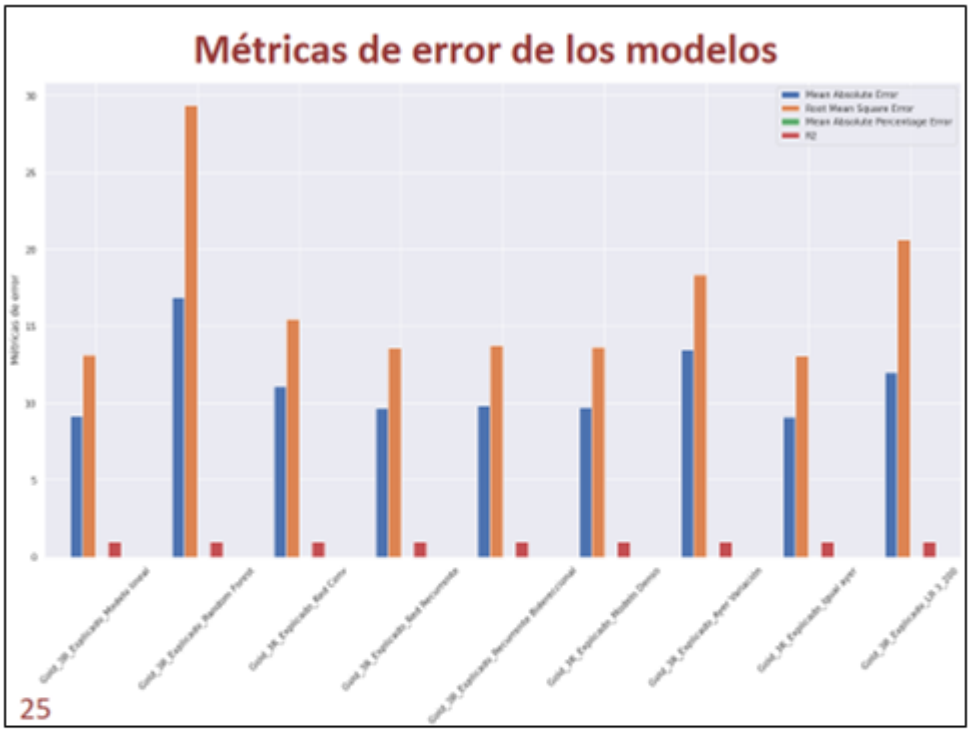


En general los algoritmos de aprendizaje automático lograron dar resultado buenos. Como es el caso de la red recurrente LSTM.
Ya veremos en la comparación de los resultados entre todos los modelos qué tan bueno fueron.

En este caso analizando el precio real vs el precio predicho se asemeja bastante a una función lineal y=x al igual que la mayoría de los modelos.
Aunque Analizando los residuos como se muestra en la fig inferior derecha estos no tienen varianza constante.



EL modelo de red recurrente bidireccional también tuvo una buena aproximación, este modelo tampoco cumple el supuesto de normalidad y homacedasticidad de los residuos al igual que los anteriores.



En este gráfico se muestra cómo se comportan los diferentes modelos utilizados.

Midiendo el error absoluto medio, la raíz del error cuadrático medio, el porcentaje del error absoluto medio y el r cuadrado de cada uno de los modelos.
Estos resultados muestran que, el modelo de regresión lineal, los modelos de redes neuronales y el modelo mañana igual que ayer se asemejan.
Por ultimo destacar que el modelo de random Forest es el que peor rendimiento tiene.

5. Experimentación y resultados

Modelo	MAE	RMSE	MAPE	R2
Mañana igual que ayer	9.128	13.119	0.006	0.998
Regresión lineal	9.150	13.148	0.006	0.998
Red neuronal densa	9.397	13.318	0.006	0.998
Red neuronal LSTM	9.683	13.602	0.007	0.997
Red neuronal BLSTM	9.773	13.708	0.007	0.997
Red Convolucional	11.129	15.465	0.008	0.997
Regresión lineal 200-1	12.020	20.625	0.008	0.994
Ayer más variación	13.483	18.354	0.009	0.995
Random Forest	16.878	29.408	0.011	0.988
Regresión lineal 30-1	48.716	85.591	0.033	0.900

26

En esta tabla se muestra los modelos ordenados de forma descendente por el error medio absoluto, que precisamente coincide también con la raíz del error cuadrático medio y El porcentaje del error medio.

Podemos ver que el modelo mañana igual que ayer es la mejor aproximación obtenida entre todos los modelos analizados hasta este momento.
Sig Diapositiva.

5. Resultados

MODELO	MAE	RMSE	MAPR	R2
Gold_1er_conjunto_datos_mañana igual que ayer	9.128	13.119	0.006	0.998
Gold_3er_conjunto_datos_Regresión lineal	9.135	13.152	0.006	0.998
Gold_4to_conjunto_datos_Regresión lineal	9.137	13.168	0.006	0.998
Gold_2do_conjunto_datos_Regresión lineal	9.141	13.112	0.006	0.998
Gold_5to_conjunto_datos_Regresión lineal	9.149	13.129	0.006	0.998
Gold_1er_conjunto_datos_Regresión lineal	9.150	13.148	0.006	0.998
Gold_1er_conjunto_datos_Red neuronal densa	9.397	13.318	0.006	0.998
Gold_4to_conjunto_datos_Red neuronal convolucional	9.495	13.552	0.007	0.997
Gold_4to_conjunto_datos_Red neuronal densa	9.507	13.501	0.007	0.998
Gold_5to_conjunto_datos_Red neuronal BLSTM	9.522	13.453	0.007	0.998

Había comentado que una de las fases en nuestra metodología sería seleccionar características, entonces se aplicaron varios métodos y a partir de esa selección de características se crean diferentes conjuntos.

En total fueron 5 conjuntos los que se utilizan, que, combinados con los 8 modelos de predicción serían 40 modelos para cada activo analizado, más dos básicos, serían un total de 42 modelos.

En la siguiente tabla se muestran los 10 mejores del total de los 42.

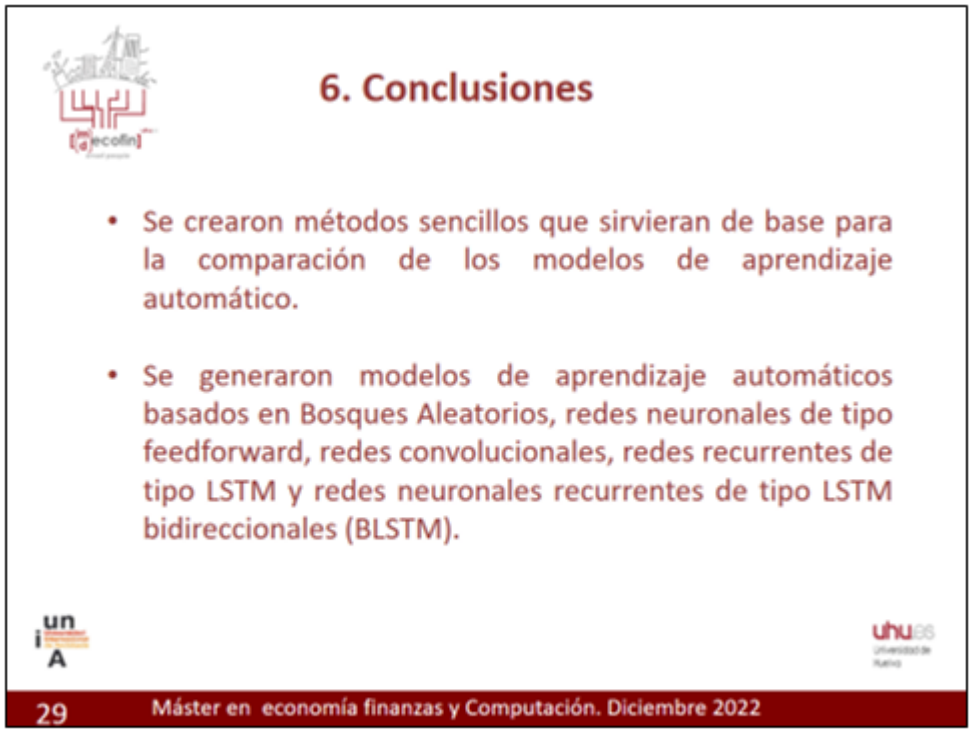
El mejor modelo sería el modelo mañana igual que ayer seguido por los modelos de regresión lineal, comentar que están los 5 modelos de regresión lineal independientemente del conjuntos de datos, luego una red neuronal densa partiendo del conjunto 1 y no es hasta la posición 10 donde se ve un modelo de red recurrente.

6. Resultados

MODELO	MAE	RMSE	MAPE	R2
BA_4to_conjunto_datos_Regresión lineal	3.419	5.408	0.016	0.996
BA_3er_conjunto_datos_Regresión lineal	3.422	5.440	0.016	0.996
BA_2do_conjunto_datos_Regresión lineal	3.422	5.432	0.016	0.996
BA_1er_conjunto_datos_mañana igual que ayer	3.426	5.430	0.016	0.996
BA_1er_conjunto_datos_Regresión lineal	3.439	5.450	0.016	0.996
BA_5to_conjunto_datos_Regresión lineal	3.441	5.449	0.016	0.996
BA_4to_conjunto_datos_Red neuronal densa	3.477	5.518	0.016	0.996
BA_2do_conjunto_datos_Red neuronal densa	3.483	5.565	0.016	0.996
BA_4to_conjunto_datos_Red neuronal convolucional	3.572	5.534	0.016	0.996
BA_4to_conjunto_datos_Regresión lineal 200-1	3.616	5.672	0.017	0.996

Mostrando los resultados sobre otro de los activos analizados como la empresa Boing Company, otras vez entre los 42 modelos analizados los mejores resultados obtenidos se dan con los modelos regresión lineal y el modelo básico mañana igual que ayer.

como conclusiones parciales se puede decir que los algoritmos de aprendizaje automático no tuvieron un mejor desempeño en la predicción de los precios de acciones y la tarea de predicción es extremadamente difícil al menos en un horizonte temporal de 24 horas como fue el caso de este estudio.



Por lo antes expuesto se plantean las siguientes conclusiones.

6. Conclusiones

- Se evaluaron los métodos, utilizando métricas de rendimiento como la raíz del error cuadrático medio, el error porcentual medio absoluto, el error medio absoluto y R cuadrado.

- Se observó, que la predicción de precios en series financieras utilizando un enfoque ingenuo y métodos clásicos de regresión lineal, pueden ser métodos efectivos.

Predicción de precios de acciones mediante modelos de aprendizaje automático

Yunio Seijo Manso

Supervisores:
Diego Marín Santos
Manuel Emilio Gegúndez Arias

Máster en economía finanzas y Computación. Diciembre 2022